Database ownership and access issues: A discussion paper

P Gioia

Department of Conservation and Land Management,
Herbarium, Kensington WA 6151
email: paulg@calm.wa.gov.au

Keywords: fauna survey, database, ownership, access

Introduction

The need to coordinate fauna survey databases across a range of private and public agencies has resulted in a proposal to house survey data in a single repository. This repository would be accessible to data contributors, such as environmental consultants, mining companies or conservation agencies undertaking fauna survey, and data users such as the Environmental Protection Authority (EPA) for the purpose of environmental impact assessment (EPA 2000), or other researchers.

The advantages of such a repository include increased consistency in survey design and consequent data capture and storage, a reduction in replication, increased collaboration between consultants and State government agencies, and an improved basis from which to understand Western Australia's ecosystems and biodiversity.

A number of important issues need to be addressed in the development of this database. Should the project be owned solely by government or a consortium of private and public participants? What access would non-contributors have? What protection or copyright issues are involved? Who would own the data and what rights would this confer? What data licensing arrangements would be required, if any? What security system should be established?

From an implementation perspective there are other important questions. What kind of model would best suit the repository – a data editing environment, a read-only warehouse or some other model? What level of ongoing data maintenance would be required and who would pay? What relationship would there be with the Western Australian Museum's specimen database, the Department of Conservation and Land Management's historical survey databases, or other related databases?

This paper canvasses some of these questions, albeit at a superficial level, for discussion purposes. It does not attempt to provided definitive statements on areas involving copyright or IP law, nor detailed specifications for a computing infrastructure. However, some recommendations are made, based on previous experience, on important requisites for an effective repository.

This paper was presented in similar form at the Workshop on the Co-ordination and Integration of Terrestrial Vertebrate Fauna Survey Database for Western Australia, held on April 5, 2002, at the University of Western Australia

Implementing a common repository

Where

A repository as described above would involve contributors from both the private and public sectors. Where should the repository be best housed – within a specific agency (be it private or public) or in a commercial facilities management situation? Factors that should be considered in this choice include longevity of the project, primary users, cost and ease of implementation and maintenance.

Data arising from fauna surveys are largely publicly funded. A common repository would include data from environmental impact assessments, biodiversity inventories and research, and research activities generally by the public arena. By its nature, a fauna survey data repository would become increasingly useful in the longer term as data are added. This would provide a broader base for data analyses and a wider scope of data to prevent the possibility of data duplication.

These arguments lend support to housing the data within a public agency having a significant legal obligation to deal with or manage conservation-related information. Agencies that fit these criteria include the Department of Conservation and Land Management, the Environmental Protection Authority and the WA Museum. The Department of Conservation and Land Management might be better suited to housing the repository because of its status as a public instrumentality and its legal obligation to manage fauna data at an operational level. Hosting by a public agency may also have lower costs than a commercial company, depending on the situation (*e.g.* where a Department is already providing a similar function). Given the public interest nature of the repository, it might be possible for a cost-effective arrangement to be worked out with the agency. That would be much less likely for a commercial company.

How

As a prerequisite, a repository needs a hosting agency with adequate existing infrastructure and commensurate biological data management skills, or the resources to provide those requirements. That infrastructure should support not only database management, but also online delivery capabilities (*i.e.* the web), and appropriate security capabilities. With respect to biological data management skills, particular regard needs to be given to the ongoing management of species-based data. The basic unit of a fauna survey is a species at a location. Depending on the fauna group, different and conflicting species classifications may exist. Within the context of

Western Australian fauna names, it is highly desirable that a single taxonomy be adopted and the custodianship of the WA Museum prevails. Furthermore, an effective fauna repository would be dependant on access to current, authoritative names from the WA Museum, provided on a regular basis.

There are also issues associated with data ageing that must be considered. Changes to species names occur as the result of taxonomic research and revision. This can be a problem in particular groups, such as invertebrates. It can be argued that tracking of names and synonymies is an essential task of any species database if it is to be of ongoing value to a broad range of users (for flora, this function is provided by the WACensus database at the WA Herbarium).

For non specimen-backed records, as is often the case with fauna survey data, name changes represent a major problem, and the integrity and usefulness of survey data will diminish over time. However, it can also be argued that any data are better than no data at all. Historical records can still be of great value to science despite problems with identification.

Ultimately, it is a decision for those implementing the repository as to what resource is applied to this problem, and what reasonable expectations users might have of the database to contain accurate, useful information for scientific research.

Consider now two approaches for how a repository could operate: implement a repository as a read-only warehouse containing a merged copy of datasets maintained elsewhere; or alternatively, develop a more traditional data-editing environment. There are pros and cons to both models, each with significant impacts on ongoing maintenance.

Read-only Warehouse

In this model, data would be maintained externally by contributors. They would be responsible for all aspects of integrity relating their own datasets. Data would be delivered in a prescribed format, adhering to specified standards as called for in EPA Position Paper 3 (EPA, 2000). Updating of species nomenclature would be the responsibility of contributors. Any changes to names would require all contributors to update their relevant datasets where required.

Those data would then be merged into the repository. If data standards have been adhered to, then the merging process should not be problematic. In this model, the repository is refreshed as often as desired and made available on a read-only basis in an agreed manner. "Home" for data resides with the contributors and corrections to data must be applied by the contributor and the repository refreshed with the new data.

The warehouse model is simple, and places the least burden on the hosting agency. Because data editing and maintenance largely occur elsewhere, a much simpler data delivery environment is required and data uploading is relatively straightforward. However, experience has shown that despite the existence of standards, there are invariably different interpretations of those standards and some data 'massaging' will be required. This arrangement places the responsibility for data maintenance on the original contributor, and because of the reality of name changes in some groups, contributors will need to provide an ongoing resource for maintaining the data. On the other hand, the contributor is like to have the greatest knowledge of the data set concerned and is therefore better placed for its ongoing maintenance.

Traditional Database Model

This model describes the way many databases are implemented; a set of data entry, validation and maintenance tools integrated with various delivery methods such as reports, online queries, *etc.* Data should, in theory, be delivered in adherence to the set standards described above. In this model, editing capabilities can be used to correct errors *in situ*, as well as enhance existing data. 'Home' for the data would become the repository. Any subsequent corrections made by contributors to source data sets would have to be carefully merged into the repository without deleting changes made therein. Alternatively, contributors would need direct editing access to the repository.

While the notion of contributor responsibility would remain intact, experience often dictates otherwise. Contributors with limited resources are more likely to provide the data as is, leaving the burden of data validation and integrity with the hosting agency. This would place greater responsibility upon the hosting agency, both in terms of increased system complexity and increased data maintenance. On the other hand, it represents an opportunity to establish a greater degree of integrity in the data.

Security

A repository as described above would have data contributed from a number of sources, both private and public sector. In some cases the data might be freely contributed; in other cases it might come as the result of conditions attached to an EIA. In some cases the data might require restricted access by virtue of confidential information on threatened species or the like, in other cases data may have access limitations imposed by custodians. Furthermore, distributed data editing access might be required, depending on which implementation approach was adopted. Thus, security mechanisms would need to be implemented which enabled arbitrary portions of the data to be screened from unauthorised access or alteration.

Custodianship and maintenance

From the above discussion, it is clear that dataset maintenance and custodianship are key issues defining the implementation of a common repository. Survey datasets are generally funded by one-off funding situations. Resources for ongoing maintenance of datasets are rarely factored into grant applications, and when they are, rarely granted. Furthermore, researchers generally do not have the resources to be encumbered by an increasing array of legacy datasets. Thus, the tendency is to relinquish responsibility for a dataset once its sponsoring project is complete and data are published. It should not be surprising, therefore, to have the idea of a common repository greeted with ambivalence by many researchers, unless it minimises their responsibility for ongoing data maintenance.

Of course, an approach could be taken whereby data are accepted as is, without further encumbrance or liability, either to the contributor or the host of the repository. This may be the only realistic approach. Users would query and interpret data on that basis. Users generally are also quick to notice inconsistencies in data. If those inconsistencies remain without any attempt at error correction, there may be less willingness to use the data meaningfully and the value of the repository may be undermined to some extent.

How much?

The initial costs of designing and implementing a repository are generally straightforward to calculate and will not be dealt with further here. It is the maintenance of the repository that that is of greater concern. If a warehouse model is adopted, data maintenance costs will be spread across contributors as the onus of data validation and integrity would fall largely on the contributor. If a data editing/value adding model is chosen, ongoing resource will be required to reconcile disparate datasets, either because of inconsistent species nomenclature or other data validation problems. In fact, both models will require a minimum level of maintenance, neither will run on their own after implementation with ongoing assistance to the hosting agency and data custodians.

Data ownership issues

The above situation might understandably create concerns for contributors regarding ownership and access issues. It is important to understand what protections are available under Federal or State law, what is capable of being owned, and what arrangements can be made to protect ownership whilst at the same time supporting the original notion of a common repository for fauna survey data.

Data in itself is not capable of being owned. The physical manifestation of the data can be owned, for example, in paper or electronic form, the manner in which data are displayed or a compilation of data may be protected as a copyright work under the Copyright Act 1968 (Commonwealth) and may also be owned. However, the most significant 'protection' which is afforded to data is the ability of the data custodian to control access to the data and through placing contractual limitations upon the use of data that is accessed.

Copyright

The following notes on copyright are by no means authoritative or exhaustive. Further information can easily be obtained from the Copyright Council website at *http://www.copyright.org.au*.

At this point in time, the only relevant form of protection for data in Australia comes from the Copyright Act 1968 (Commonwealth). Copyright law does not protect ideas or information, and raw data is in itself *not* protected by copyright. However, the expression of that data can be, including compilations of data, so long as it meets certain criteria. A compilation may be protected by copyright if "there has been sufficient intellectual effort in the selection or arrangement of the information; or if the author has engaged in sufficient work or incurred sufficient expense in gathering the information, even if there is no creativity involved in the selection or arrangement" (ACC, 2001a). The definition of what constitutes "sufficient" is obviously open to argument. Thus, so long as there is sufficient effort incurred in a particular database by its author, the database is protected by copyright. However, there is nothing to prevent someone using the same source data to produce another compilation, so long as they meet the above criteria. According to copyright law, even forms and tables can be copyright, so long as tests of originality are met.

Copyright in databases or compilations will only be infringed if the alleged infringer copies a substantial portion of the work. Extraction of a single or a small number of pieces of data will not infringe the copyright, neither will the extract of a large amount of data if it is reproduced in a different form which is not a substantial reproduction of the original database or compilation.

Normally, the author is the copyright owner, with certain exceptions. These exceptions include employees where the work is undertaken as part of normal duties for an employer. A State, Territory or Federal Government is normally the first owner of copyright for data assembled by, or for, public servants. In the case of a consultant it depends on the terms of the consultancy. Furthermore, copyright can be both assigned or licensed via written agreement (ACC, 2001c). Interestingly, under the Copyright Act, governments, including State Governments and qualifying agencies, can copy any copyright material without infringing copyright, so long as the copying of material is for government purposes having a demonstrable public interest component. Use of copyright material may require notification of the copyright owner and possibly a written agreement. A fee may be required to be paid to the copyright owner.

Government departments have no independent legal status. The copyright of material created by, or for, government departments and agencies is owned by the relevant State or Commonwealth Crown. Thus, issues of access to data between and internal to Government agencies are not determined by ownership, but administrative policy. More general details on government and copyright can be obtained from the Australian Copyright Council (ACC, 2001b).

Implications for a common fauna survey data repository and data licensing

If, as is suggested above, a repository is implemented and/or hosted by a State agency, access to the data compiled therein would be controlled by the hosting agency on behalf of the State. This would be the case irrespective of who provided the raw data. The terms of access could be modified by a data licensing agreement.

For a common fauna repository to work effectively, a suitable data licensing agreement would be required. A key issue the license should define is identification of parties to the licence (*ie* who's in the "club"). The license should anticipate the kinds of contributors to, or users of the repository, and purposes for which the data may be used, particularly in prospective commercial

environments. In other words the license should define who a participant in the repository is.

Other areas the license should deal with include:

- acceptable purposes for use of data,
- responsibilities of the hosting agency in maintaining database infrastructure and ensuring access to license partners,
- responsibilities of contributors in providing data to an agreed standard with agreed levels of ongoing maintenance,
- determination of costs associated with the maintenance and provision of data,
- indemnity issues,
- termination,
- data transfer to non-participants, and
- other specific conditions.

Examples of licensing agreements are available on the web (*e.g.* WALIS, 2000) or on request from agencies such as the Department of Conservation and Land Management, any of which could be used as a starting point. It is likely that there would be a number of different licenses to cover specific situations, including both commercial and non-commercial situations.

Furthermore, by developing a *Memorandum of Understanding* (MOU) between all parties, the licensing agreement could be substantially simplified.

Conclusions

The implementation of a fauna survey repository accessible by a wide variety of users for the purposes of environmental impact assessment or further research is feasible. However, such a repository would be best implemented within a State agency having an ongoing responsibility in the management of fauna data. That agency should also have adequate resources, not just for hosting the repository, but its implementation phase and ongoing maintenance.

There would need to be clear lines of responsibility for the ongoing maintenance of data, particularly with regard to species names, and a clear data standard to adhere to.

A suitable data licensing agreement and/or MOU is essential, to which all contributors or users would be party to, so that future terms of access are clearly understood and agreed to.

Acknowledgements: Thanks to Peter Van Bruchem from the Intellectual Property Support Program, Department of Industry and Technology, and Robin Piesse from the WA Land Information System Program, for their helpful comments in reviewing the paper.

References

ACC (2001a) Compilations, tables and forms: copyright protection. Information sheet G66v02. Australian Copyright Council, July, 2001, Redfern.

ACC (2001b) Copyright in Australia: an introduction. Information sheet G10. Australian Copyright Council, March, 2001, Redfern.

ACC (2001c) Government and copyright. Information sheet G62. Australian Copyright Council, August, 2001, Redfern.

DIT (2000) Intellectual Property Guidelines. Third Edition (Draft). Department of Industry and Technology, October, 2000, Perth.

EPA (2000) General requirements for terrestrial biological surveys for environmental impact assessment in Western Australia. Position paper No. 3 (preliminary). Environmental Protection Authority, May, 2000, Perth.

WALIS (2000) WALIS Licensing Agreements. Western Australian Land Information System, June, 2000. URL: *http://www.walis.wa.gov.au/walis/content/licensing_agreements.html*